Plague Dot Text

Journal of Data Mining and Digital Humanities

Plague Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894-1952)

Arlene Casey[1], **Mike Bennett**[2], **Richard Tobin**[1], **Claire Grover**[1], **Iona Walker**[3], **Lukas Engelmann**[3], **Beatrice Alex**[1,4]

[1]Institute for Language, Cognition and Computation, School of Informatics
[2]Digital Library Team, University of Edinburgh Library
[3]Science Technology and Innovation Studies, School of Social and Political Science
[4]Edinburgh Futures Institute, School of Literatures, Languages and Cultures

University of Edinburgh, Edinburgh, UK

Corresponding author: Beatrice Alex , balex AT ed.ac.uk

Abstract

The design of models that govern diseases in population is commonly built on information and data gathered from past outbreaks. However, epidemic outbreaks are never captured in statistical data alone but are communicated by narratives, supported by empirical observations. Outbreak reports discuss correlations between populations, locations and the disease to infer insights into causes, vectors and potential interventions. The problem with these narratives is usually the lack of consistent structure or strong conventions, which prohibit their formal analysis in larger corpora. Our interdisciplinary research investigates more than 100 reports from the third plague pandemic (1894-1952) evaluating ways of building a corpus to extract and structure this narrative information through text mining and manual annotation. In this paper we discuss the progress of our ongoing exploratory project, how we enhance optical character recognition (OCR) methods to improve text capture, our approach to structure the narratives and identify relevant entities in the reports. The structured corpus is made available via Solr enabling search and analysis across the whole collection for future research dedicated, for example, to the identification of concepts. We show preliminary visualisations of the characteristics of causation and differences with respect to gender as a result of syntactic-category-dependent corpus statistics. Our goal is to develop structured accounts of some of the most significant concepts that were used to understand the epidemiology of the third plague pandemic around the globe. The corpus enables researchers to analyse the reports collectively allowing for deep insights into the global epidemiological consideration of plague in the early twentieth century.

Keywords

historical text mining, annotation, corpus access and analysis, the third plague pandemic

INTRODUCTION

The Third Plague Pandemic (1894 - 1950), usually attributed to the outbreak in Hong Kong in 1894, spread along sea trade routes affecting almost every port in the world and almost all inhabited countries, killing millions of people in the late nineteenth and early twentieth centuries [Engelmann, 2018, Echenberg, 2007]. However, as outbreaks differed in severity, mortality and longevity, questions emerged at the time of how to identify the common drivers of the epidemic. After the Pasteurian Alexandre Yersin had successfully identified the epidemic's pathogen in

1894, *yersinia pestis* [Yersin, 1894], the attention of epidemiologists and medical officers turned to the specific local conditions to understand the circumstances by which the presence of plague bacteria turned into an epidemic. These observations were regularly transferred into reports, written to deliver a comprehensive account of the aspects deemed important by the respective author. These reports included discussions, such as extensive elaborations on the social structure of populations, long descriptions of the built environment or close comparison of the outbreak patterns of plague in rats and humans. Many of the reports were quickly circulated globally and served to discuss and compare the underlying patterns and characteristics of a plague outbreak more generally.

These reports are the underlying data set for ongoing work in the *Plague.TXT* project which is conducted by an interdisciplinary team of medical historians, computer scientists and computational linguists. While each historical report was written as a stand-alone document relating to the spread of disease in a particular city, the goal of our work is to bring these reports together as one systematically structured collection of epidemiological reasoning about the third plague pandemic. Given that most reports are already under public domain, this corpus can then be made available to the wider research community through a search interface. Using methodology from genre analysis [Swales, 1990], our approach looks to identify common themes used in the narrative to discuss aspects, such as *conditions, treatments, causes, outbreak history*. These themes will then be linked across the report collection. This allows for comparative analysis across the collection e.g. comparing discussion on *treatments* or *local conditions*. In addition to structuring the narrative by theme we also annotate the collection for entities, such as *dates, locations, distances, plague terms*. This provides for a rich source of information to be tracked and analysed throughout the collection which may unveil interesting patterns with regard to the spread and interpretation of this pandemic.

In the following sections we give an overview of our pilot study firstly describing the background for this project, the data collection, the challenges presented by OCR and improvements we have made to the original digitised reports. Following this we describe our annotation process including our model to structure the reports to extract information. We discuss our combination of manual annotation and automated text mining techniques that support the retrieval and structuring of information from the reports. We discuss our search interface, enabled through Solr, which we use to make the collection available online. Finally, we give some examples of potential use cases for this interface.

I BACKGROUND AND RELATED WORK

The report collection used in *Plague.TXT* project is a valuable source for multiple historical questions. The pandemic reports offer deep insights into the ways in which epidemiological knowledge about plague was articulated at the time of the pandemic. While they often contain a wealth of statistics and tabulated data, their main value is found in articulated viewpoints about the causes for a plague epidemic, about the attribution of responsibility to populations, locations or climate conditions as well as about evaluating various measurements of control.

Analysis of reports of the third plague pandemic have been conducted previously, although these centre mainly on manual collation of data using quantitative methods, such as collecting statistics across reports for mortality rates. This derived data has been used to reconstruct transmission trees from localised outbreaks [Dean et al., 2019], and to study potential sources and transmission across Europe [Bramanti et al., 2019]. Our *Plague.TXT* project moves beyond existing work by aiming to digitally map epidemiological concepts and themes from the collection of reports, developing pathways to extracting quantitative as well as qualitative in-

formation semi-automatically. Combining text mining and manual annotation, we seek to analyse historical plague reports with respect to their narrative structure. This allows us to collate section-specific information, e.g. *treatments* or discussions on *causes*, for analysis and research.

From the perspective of historiography, this approach also encourages systematical reflections on the underlying conventions of epidemiological writing in the late nineteenth and early twentieth century. Rather than considering reports only within their specific local and historical context, the lateral analysis outlined below contributes to a better understanding of the history of epidemiology as a narrative science [Morgan and Wise, 2017]. As we engage with the ways in which epidemiologists argued about outbreaks, we identify the concepts they used to investigate the same disease in different locations. This lateral approach contributes to a better understanding of how these reports were conceived with reference to descriptions and theories used in other reports, and to understanding the epistemological conditions under which epidemiological knowledge began to be formalised at the time on a global scale [Morabia, 2004, Engelmann, 2018].

1.1 Challenges of Understanding Historical Text with Modern Text Mining Tools

Whilst there has been a wealth of new tools produced within the text mining community in recent decades, they cannot always be directly used with historical text requiring adaptation for historical corpora. These changes are due to language evolution resulting not only in differences in style, but also in aspects such as, vocabulary, semantics, morphology, syntax and spelling. Spelling in historical texts is known to exhibit diachronic, changes over time, and synchronic variance, inconsistencies within the same time period, due to, for example, differences in dialect or spelling conventions [Piotrowski, 2012]. There have been numerous approaches to spelling normalisations, such as those based on rules or edit distances [Bollmann, 2012, Pettersson et al., 2013a, Mitankin et al., 2014, Pettersson et al., 2014], statistical machine translation [Pettersson et al., 2013b, Scherrer and Erjavec, 2013] and more recently neural models [Bollmann et al., 2017, Korchagina, 2017]. The process of OCR translation of the historical text itself can bring spelling irregularities. We account for spelling problems in three ways in this work, (cf. Sections 3.2, 4.1). Firstly, we use a lexicon which contains spelling variants found during the pilot annotation when automatically tagging entities, secondly, during the manual annotations any entities have correct spellings entered. Finally, when inputting our corpus to Solr we correct for spelling mistakes. We say more about general OCR issues in historical text and how we make improvements in processing the OCR generated for our corpus in Section 2.1.

Differences in syntax can also lead to challenges in using existing NLP tools, such as named entity taggers or part-of-speech taggers [Thompson et al., 2016]. These rely on accurate identification of syntactic relationships, such as sequences of nouns and adjectives, and word order can be stricter in modern day languages [Campbell, 2013, Ringe and Eska, 2013]. Changes of semantic meaning of words provide further challenges, e.g. widening and narrowing of word senses or change in terminology over time. These can cause issues, such as a reader today may interpret the meaning differently to how it would commonly have been interpreted at the time [Pettersson, 2016]. This could also cause problems when searching historical text when different terminology is used for searching, resulting in no results or results that are hard to interpret for the modern-day user. Using the functionality within Solr we intend to create a map of any such semantic changes to support the challenges that users may encounter when searching.

1.2 Understanding the Genre of Outbreak Reports

In the reports, outbreaks were described by their occurrence over time. Statistical data was often made meaningful through descriptions and theoretical explorations. If the authors attributed

causality, they commonly presented these through careful deliberation of often contradicting hypotheses and theories. Some authors structure their reports using an *introduction, outbreak history, local* or *geographical conditions* followed by discussion on *causes, treatments* and then perhaps a section on *cases*. Other reports combine this information into sections that discuss all these aspects about a specific location or town and then progress onto a similar discussion about the next town and its localised outbreak. Identifying the narrative structure is challenging as each report differs in its presentation, ordering and style of content.

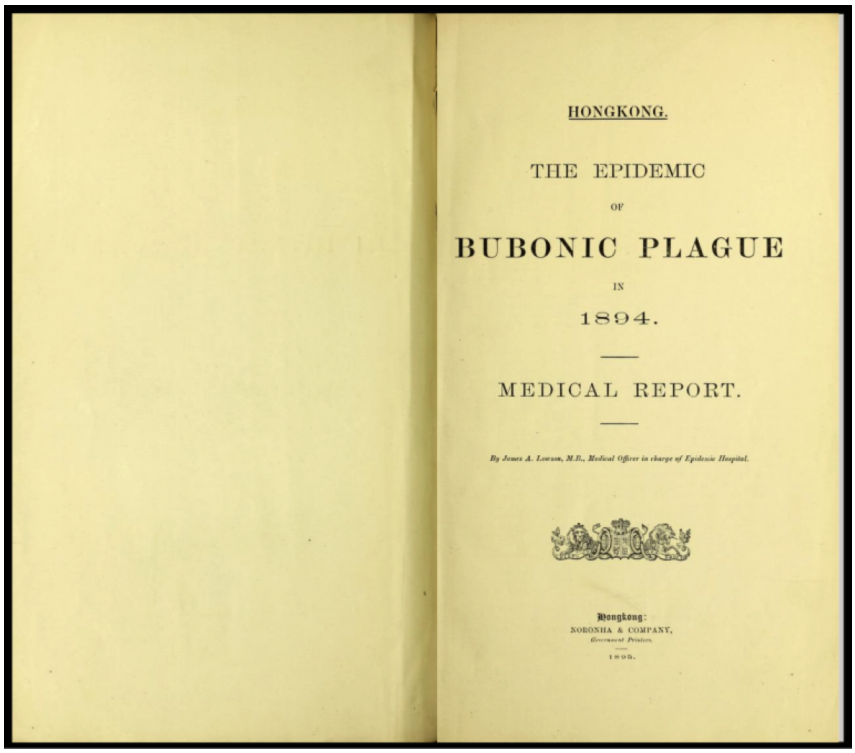
Despite their differences in style, these communicative reports are intended for the same audience of government officials and fellow epidemiologists and will present their arguments comparably. This study of discourse that shares communicative purpose is called genre analysis [Swales, 1990]. Recent decades have seen considerable contributions to understanding how authors structure arguments within specific genres and have demonstrated pathways of how text mining can be applied to automatically recognise these structures. Most relevant to our work is that done for scientific articles, such as Argument Zoning [Teufel, 1999] or Core Scientific Concepts [Liakata et al., 2012]. These works seek to model the intentional structure of a research article. However, the models proposed are designed to extract different information from different disciplines. For example, Argument Zoning is originally designed for use within the discipline of Computational Linguistics. When this model is applied within the domain of Chemistry [Teufel et al., 2009] it is required to extend it to adequately address aspects of communication that occur within this domain. Although other work that models communicative purpose may have similar goals to ours, it does not adequately represent our needs and does not capture all the aspects of our type of discourse. Hence, we had to develop our own model to capture the information and arguments made within our collection of reports.

Using genre analysis as our methodological approach, treating each report as a communicative event about a specific plague outbreak, our focus is on building a structure to label the information contained in individual reports, such that similar discourse segments can be linked and studied across the collection. We seek to collate the concepts, themes and approaches across the report collection to consider comparable conventions within the entire corpus. For example, we seek to enable comparative analysis of *causes, treatments, local conditions* between different outbreaks, and various times and places. While there has been previous work on bootstrapping and mapping concepts in other types of historical texts (e.g. commodities in historical collections on nineteenth century trade in the British empire [Klein et al., 2014, Hinrichs et al., 2015, Clifford et al., 2016]) we are unaware of other work where this is done with respect to narrative document structure across a historical collection.

1.3 Contribution

Our contribution is the development of a systematically structured corpus, which we capture through annotation, to assimilate similar discourse segments such as *causes* or *treatments* across the reports. In addition, we develop an interactive search interface to our collection.

This search tool in combination with our structure model allows follow-on research to conduct automated or semi-automated exploration of a rich source about the conceptual thinking at the time of the third plague pandemic. This allows for better understanding of the historical epistemology of epidemiology and to thus provide valuable lessons about dealing with contemporary global spread of disease. The corpus thus constitutes an archive, from which future analysis will discern *concepts*, with which plague has been shaped into an object of knowledge in modern epidemiology. This will also enable new perspectives on the formalisation of epidemiology as a discipline in the twentieth century.


HONGKONG.

THE EPIDEMIC

OF

BUBONIC PLAGUE

IN

1894.

MEDICAL REPORT.

By James A. Lowson, M.B., Medical Officer in charge of Epidemic Hospital.

Hongkong:
NORONHA & COMPANY,
Government Printers.
1895.

Figure 1: Bubonic plague report for Hong Kong [Lowson, 1895].

II DATA

The third plague pandemic was documented in over 100 outbreak reports for most major cities around the world. Many of them have been digitised, converted to text via OCR and are available via the Internet Archive[1] and the UK Medical Heritage Library.[2] Figure 1 shows an example of such a report covering the Hong Kong outbreak which was published in 1895 and is accessible with open access on Internet Archive.

We treat all relevant reports for which we have a scan as one collection. While the majority of reports in this set (102) are written in English, there are further reports in French, Spanish, Portuguese and other languages which we excluded from the analysis at this stage.

The years of publication of English reports in the collection are visualised in the histogram shown in Figure 2 grouped by 5-year intervals. The majority of reports were published a few years after the plague pandemic started between 1895 and 1915. A few more reports were published during the tail-end of the pandemic leading up to 1950. The pandemic was not officially declared over by the World Health Organisation until 1960 when the number of cases dropped below 200 worldwide. However, our collection does not include any reports beyond 1950 as after then there were no major significant outbreaks.

The main locations of the outbreaks described in the reports are visualised in Figure 3. The size of each mapped location corresponds to the number of reports covering it. Most reports are about San Francisco, Hong Kong and Bombay but there is a long tail of less frequently covered locations. The map shows that many of them are located along the coast, cities with ports where the plague spread particularly easily as a result of ongoing trade at that time. Some locations

[1] https://archive.org/details/b24398287

[2] https://wellcomelibrary.org/collections/digital-collections/uk-medical-heritage-library/

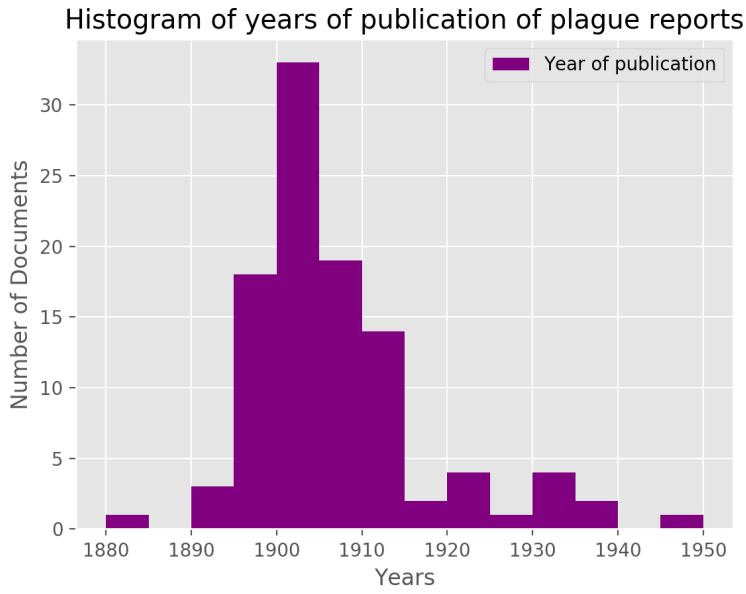


Figure 2: Histogram of years of publication for the 102 English reports in the collection.

are inland and correspond to country or region names with corresponding latitude and longitude coordinates retrieved from GeoNames.[3]

Counts	Total	Min	Max	Mean	Stddev
Sentences	229,043	32	17,635	2,245.5	3,713.6
Words	4,443,485	1,091	396,898	43,563.6	74,621.0

Table 1: Number of sentences and words in the collection of English plague reports, as well as corresponding counts for the smallest document (Min) and the largest document (Max), the average (Mean) and standard deviation (Stddev).

Table 1 provides an overview of counts of sentences and words in the collection and illustrates the variety of documents in this data. To derive these counts we used automatic tokenisation and sentence detection over the raw OCR output which is part of the text mining pipeline described in section III. While the smallest document is only 32 sentences long containing 1,091 word tokens, the largest report contains almost 400,000 word tokens. The collection contains 38 documents with up to 5,000 words each, 15 reports with between 5,000 and 10,000 words each, 32 documents with between 10,000 and 100K words each and 17 documents with 100K or more words each. In total, the reports amount to over 4.4 million word tokens and over 229,000 sentences.

Exact details on what articles or works are part of this collection and accessible download links to their pdfs (if available) are provided on the project's GitHub repository.[4]

2.1 OCR Improvements

When initially inspecting this digitised historical data, we realised that some of the OCR was of inadequate quality. We therefore spent time during the first part of the project on improving the OCR quality of the reports.

[3] https://www.geonames.org
[4] https://github.com/Edinburgh-LTG/PlagueDotTxt

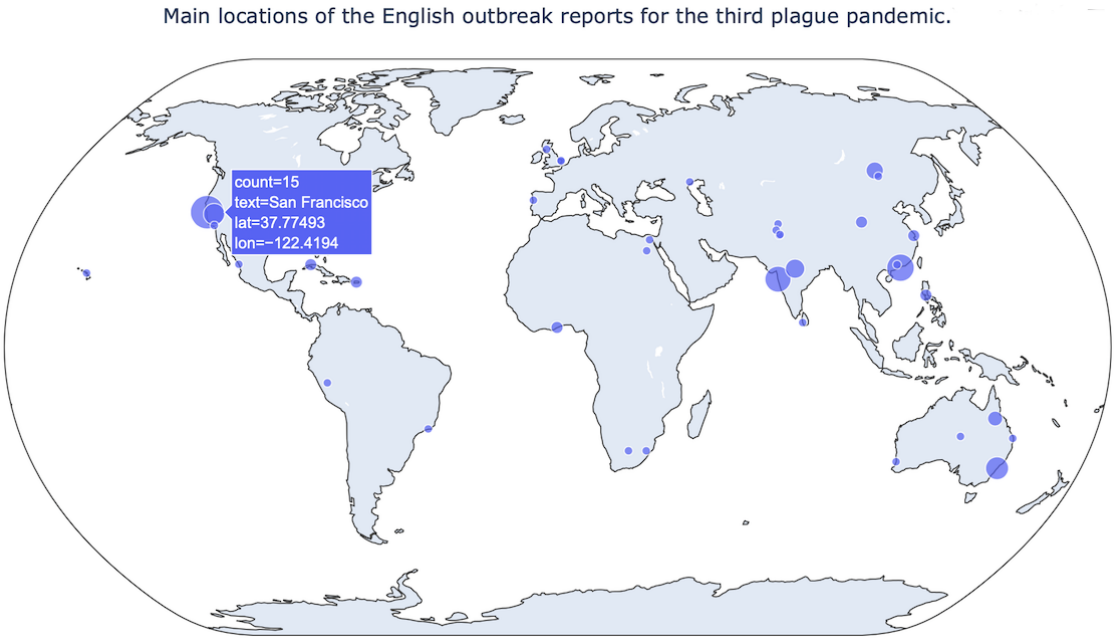


Figure 3: Geographical scatter plot of the main outbreak locations of the English reports in the dataset.

Using computer vision techniques, we processed the report images to remove warping artefacts [Fu et al., 2007]. This was done using Python and the numpy,[5] SciPy,[6] and OpenCV[7] libraries. We find the text within each image by binarising and thresholding the image,[8] followed by horizontal dilation to connect adjacent letters. Following this, principal component analysis was used to determine the location of text lines in the image, and then OpenCV is used to estimate the "pose" of the page and generate a reprojection matrix, which is optimised with SciPy using the Powell solver, an optimisation algorithm available in this library.[9] An example page image before and after dewarping is shown in Figure 4.

We then identified likely textual areas in report images, and produced an effective crop, to provide the OCR engine with less extraneous data (see Figure 5). This was done with similar methods to the page dewarping, again utilising the OpenCV library to binarise, threshold and dilate the text components of the image. This process was repeated until a maximum target number of contours were present in the image, and then a subset-sum was used to find the most efficient crop. More information on the methods used and steps taken can be found in the University of Edinburgh Library Labs blog post.[10] OCR was then performed using Tesseract,[11] trained specifically for typeface styles and document layouts common to the time period of the reports.

Training was done across a range of truth data, covering period documents obtained from the

[5] https://numpy.org
[6] https://www.scipy.org
[7] https://opencv.org
[8] We applied an adaptive 65% threshold which helped to preserve the text on the page and remove blemishes and text bleeding from the printing on the reverse of a page.
[9] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-powell.html
[10] http://libraryblogs.is.ed.ac.uk/librarylabs/2017/06/23/automated-item-data-extraction/
[11] https://opensource.google.com/projects/tesseract

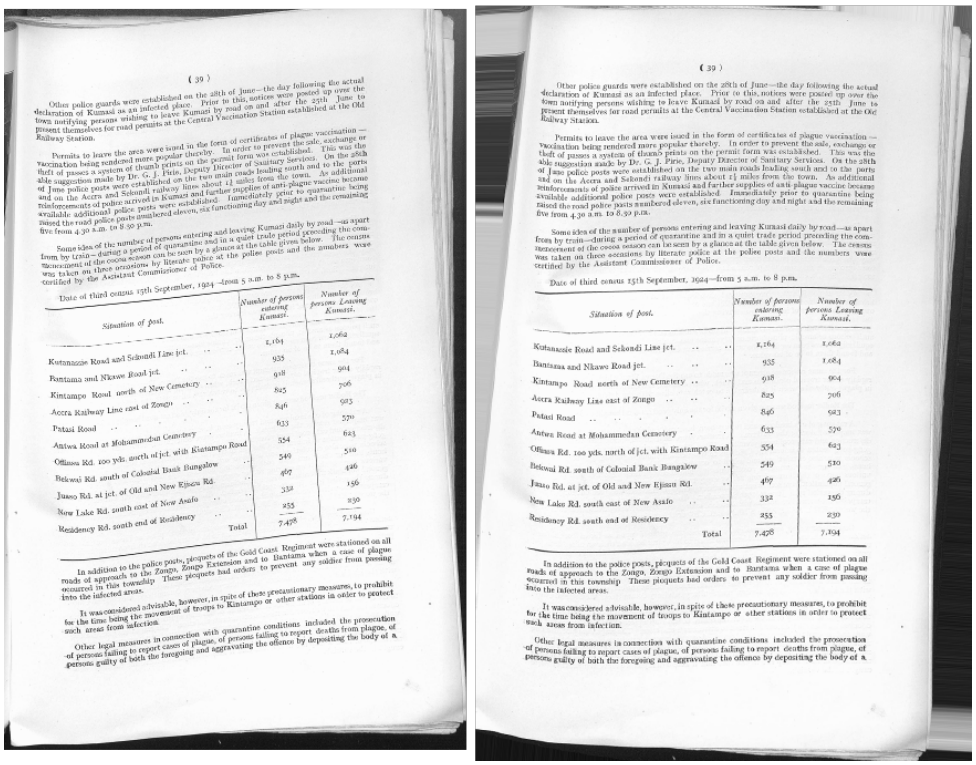


Figure 4: A sample page from one of the reports before and after the dewarping process.

IMPACT Project data sets,[12] documents from Project Gutenberg prepared for OCR training,[13] internal ground-truth data compiled as part of the Scottish Session Papers project at the University of Edinburgh[14] and typeface data sets designed for Digital Humanities collections.[15]

While we have not yet formally evaluated the improvements made to the OCR, observation of the new OCR output shows clear improvements in text quality.[16] This is important as it affects the quality of downstream text mining steps. Previous research and experiments have found that errors in OCRed text have a negative cascading effect on natural language processing or information retrieval tasks [Hauser et al., 2007, Lopresti, 2008, Gotscharek et al., 2011, Alex and Burns, 2014]. In future work, we would like to conduct a formal evaluation comparing the two versions of OCRed text to quantify the quality improvement.

Figures 6 and 7 shows a comparison of the OCR for two excerpts from the Hong Kong plague report referred to earlier [Lowson, 1895]. The excerpts marked as "Available OCR" refer to the version openly accessible on Internet Archive and created using ABBYY FineReader 11.0.[17] The improved OCR was created using Tesseract as part of the work presented in this paper. Errors in the OCR are highlighted in red. While a thorough evaluation across the OCRed reports in the corpus is needed to provide a quantitative comparison of OCR quality for both methods, these example excerpts illustrate the types of errors created by them. Initial observations suggest

[12] https://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/

[13] https://github.com/PedroBarcha/old-books-dataset

[14] https://www.projects.ed.ac.uk/project/luc020/brief/overview

[15] https://github.com/jbest/typeface-corpus

[16] Previous work conducted by members of the IMPACT project has formally compared OCR quality of ABBYY FineReader and Tesseract [Heliński et al., 2012] for different types of test sets. They have shown that the latter performs more accurately on gothic type documents in terms of both character and word level accuracy.

[17] https://www.abbyy.com/media/2761/abbyy-finereader-11-users-guide.pdf

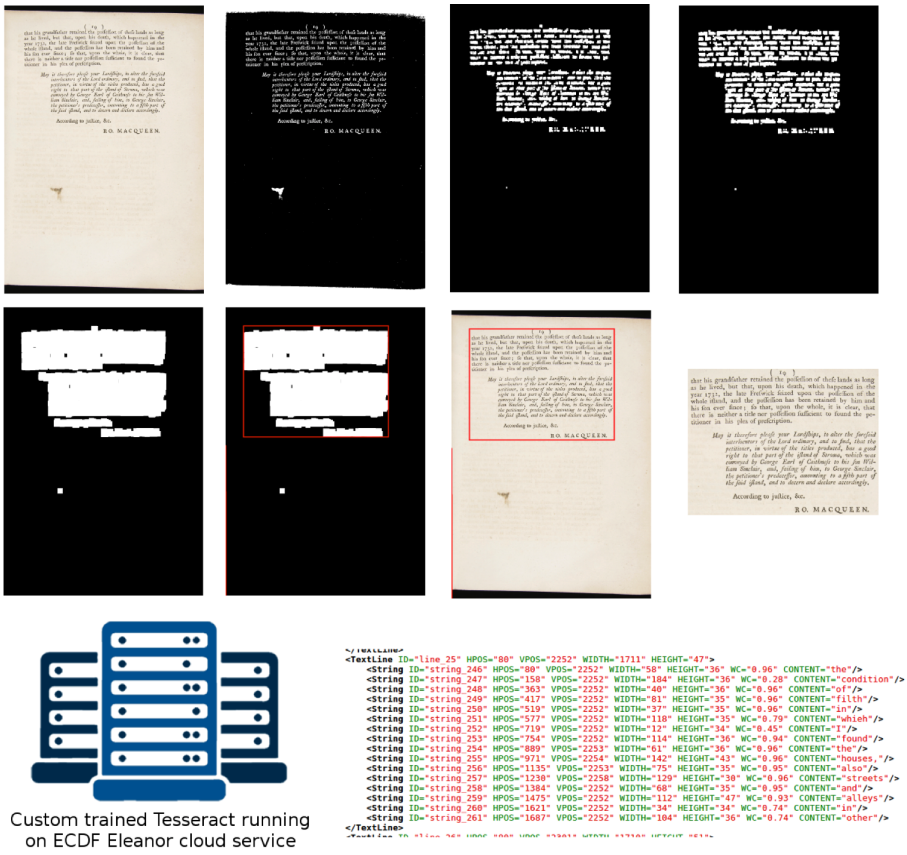


Figure 5: The OCR process.

Available OCR

I have tlie lioiiour to foi'ward herewith foi- your information a Heport upon the Kpideiiiic of Plague in HoDgkoug in 1894, so far as it concerns tlie medical work which I carried out under your directions. I regret extremely that several important matters — including the e[)idemiology of the disease — which I could have wished to discuss at some length, have been touched upon very superficially, or passed over altogether in this Report. I will ask you to accept as an excuse for my shortcomings in these respects the following facts of which you are, I believe, already cognizant:

Improved OCR

J have the honour to forward herewith for your information a Report upon the Epidemic of Plague in Hongkong in 1894, so far as it concerns the medical work which I carried ont under your directions. I regret extremely that several important matters — including the epidemiology of the disease — which I could have wished to discuss at some length, have been touched upon very superficially, or passed over altogether in this Report [will ask you to accept as an exeuse for my shortcomings in these respects the following facts of which you are, I believe, already cognizant:

Figure 6: Excerpt from the Hong Kong report with different versions of OCR output. The Internet Archive image containing this excerpt can be accessed here: https://archive.org/details/b24398287/page/n7

Available OCR

That the Litrines are a source of))ropao^atino: the infection as described by Dr. LowsoN there is no doubt, and |)roc)f is afforded by the dates of the closiiiii; of the surrounding liouses. I found on inquiry that during- the end of jMay and the beginning of June, when the jn-evailing winds were from the east and north, the houses to the west and south of the latrines were closed and afterwards, when the prevailing winds were fi'om the south and west, the houses to the north and east of the latrines were closed, being fonnd infected and more than three deaths having occurred in each of them. Mr. RA^[made elaborate phins of the City of Victoria showing where the plague existed, and the proportion of houses in each district that were infected.

Improved OCR

That the latrines are a source of propagating the infection as described by Dr. Lowson there is no doubt, and proof is afforded by the dates of the closing of the surrounding houses. | Found on inquiry that during the end of May «and the beginning of June, when the prevailing winds were from the cast and north, the houses to the west and south of the latrines were closed and afterwards, when the prevailing winds were from the south and west, the houses to the north aud east of the latrines were closed, being found infected and more than three deaths having occurred in each of them. Mr. Ram made elaborate plans of the City of Victoria showing where the plague existed, and the proportion of houses in each district that were infected,

Figure 7: Another excerpt from the Hong Kong report with different versions of OCR output. The Internet Archive image containing this excerpt can be accessed here: https://archive.org/details/b24398287/page/n7

that the first method appears to struggle to recognise common words like *honour* or *latrines* and names like *Hongkong* and *Dr. Lowson* correctly. The Tesseract model appears to be more robust towards names and common words in these examples but, in contrast, makes mistakes for the personal pronoun *I* and function words like *and* or *out*. As our analysis is primarily focused on content words, observing the output led us to choose the results produced by the Tesseract model for further processing and annotation.

III ANNOTATION

This section describes the schema we implemented to structure the information contained within the reports and automatic and manual annotation applied to our collection of plague reports. We first processed them using a text mining pipeline which we adapted and enhanced specifically for this data. The text mining output annotations are then corrected and enriched during a manual annotation phase which is still ongoing. Each report that has undergone manual annotation is then processed further using automatic geo-resolution and date normalisation.

3.1 Developing an Annotation Schema

As discussed in the Background section, our methodological approach is based on genre analysis [Swales, 1990, Bhatia, 2014] which treats each report as a communicative event. We hypothesise that the reports - despite their variation of styles - will present and structure their arguments comparably, as they are intended for the same audience of fellow epidemiologists and government officials. Thus we assume to find comparative segments of text which discuss a similar theme e.g. *measures taken, local conditions* across the collection of reports. We refer to these comparative segments of text as zones where each zone has a specific purpose, described in Table 2. Within each zone, the author uses the narrative to build an argument or convey thinking about the zone's theme. For example, within a *measures* zone, authors have discussed measures taken to prevent the spread of the disease and their impact. The collation of report narratives into zones is not straightforward as authors approach the narrative with dif-

ferent styles and label text with different titles. For example, one may call a section of text *Background* and another may call it *Outbreak History* making the application of a schema to support automated labelling challenging. Therefore labelling of zones is done manually by annotating text through close reading of the report. However, in the future we intend to investigate if this could be approached in an automated way. The list of zones we have chosen as a scheme for annotation has emerged both from the formal conventions of published reports (with regards to the report's apparatus, containing *title-matter, preface, footnotes*) as well as from extensive historical research. Zones that emerged from sections and chapters within some reports were aligned with overarching concepts and categories, which epidemiologists used at the time.

In addition to our zoning schema, we also annotated for a number of entities within the text (Table 3). This supports the comparative analysis across zones allowing entities to be tracked, such as *location*, *plague term*, *date* and *time*. The zoning schema and entity list was created from studying a subsection of reports, section titles and three rounds of pilot annotation on a subsection of documents.

3.2 Automatic Annotation and Text Mining

To process the plague reports, we used the Edinburgh Geoparser [Grover et al., 2010], a text mining pipeline which has been previously applied to other types of historical text [Rupp et al., 2013, Alex et al., 2015, Clifford et al., 2016, Rayson et al., 2017, Alex et al., 2019, for example]. This tool is made up of a series of processing components. It takes as input raw text and performs standard text pre-processing on documents in XML format, including tokenisation, sentence detection, lemmatisation, part-of-speech tagging and chunking as well as named entity recognition and entity normalisation of dates and geo-resolution in the case of location names (see Figure 8). The processing steps are applied using LT-XML2, our in-house XML tools [Grover and Tobin, 2006].[18] Before tokenisation, we also run a script to repair broken words which were split in the input text as a result of end-of-line hyphenation described in [Alex et al., 2012].

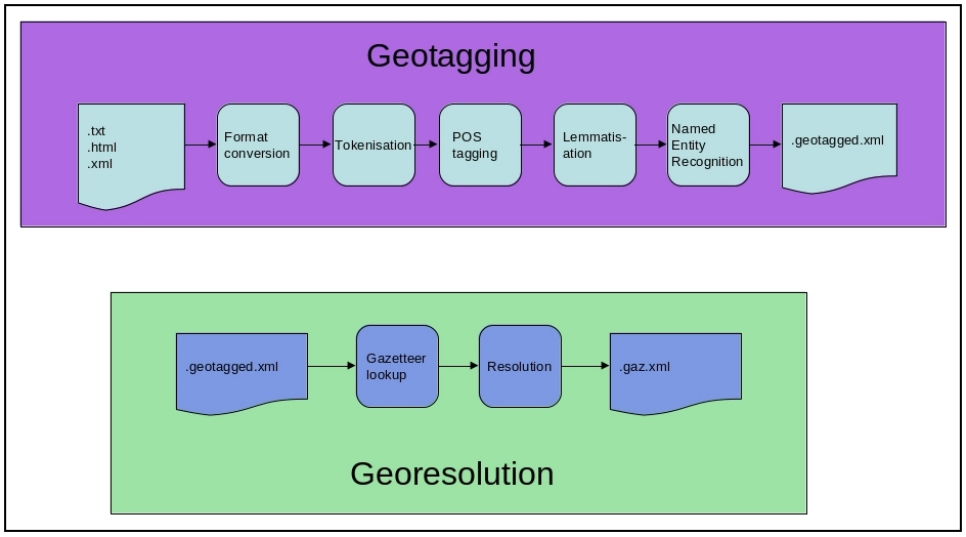


Figure 8: The Edinburgh Geoparser pipeline.

We adapted the Edinburgh Geoparser by expanding the list of types of entities it recognises in text, including geographic-feature, plague-ontology-term and population/group of people etc.[19]

[18] https://www.ltg.ed.ac.uk/software/ltxml2/

[19] Note that our goal was to emulate descriptions used by the authors at the time, mirroring concepts of race

Zones	Description
Title-matter	Title page
Preface	Preface information
Content-page	Content page information
Introduction	State of the epidemic at the time of the production of the report, summary of key features, evaluation of significance of the epidemic
Disease history	General points on the history of the epidemic, origin of outbreak
Outbreak history	Geographical and chronological overview of local outbreak. What happened this place this year
Local conditions	Descriptions of key elements that are considered noteworthy, something that has contributed or impacted the outbreak
Causes	Causes identified by the author e.g. usually points of origin, specific local conditions or descriptions of import
Measures	List of the measures e.g. undertaken to curb the outbreak, sanitary improvements, quarantines, disinfection or fumigation and rat catching
Clinical appearances	Description of the disease appearance, its usual course and its mortality
Laboratory	Description of bacteriological analysis, human lab work
Treatment	Description of the treatment given to patients
Cases	List of individual cases, usually with age, gender, occupation, course of disease, and time and dates of infection and death
Statistics	Contains many lists or tables of statistics such as deaths
Epizootics	Contains information solely about animals, experiments or discussions
Appendix	Labelled appendix
Conclusion	Conclusion

Table 2: Zones

Entity Type	Entity Mentions
person	Professor Zabolotny, Professor Kitasato, Dr. Yersin, M. Haffkine
location	India, Bombay, City of Bombay, San Francisco, Venice
geographic-feature	house, hospital, port, store, street
plague-ontology-term	plague, bubo, bacilli, pneumonia, hemorrhages, vomiting
date	1898, March 1897, 4th February 1897, the beginning of June, next day
date-range	1900-1907, July 1898 to March 1899, since September 1896
time	midnight, noon, 8 a.m., 4:30 p.m.
duration	ten days, months, a week, 48 hours, winter, a long time
distance	20 miles, 100 yards, six miles, 30 feet
population/group of people	Chinese, Europeans, Indian, Russian, Asiatics, coolies, villagers
percent	8%, 25 per cent, ten per cent

Table 3: Entity types and examples of entity mentions in the plague reports.

A list of entity types extracted from the plague reports and examples are presented in Table 3. Date entity normalisation and geo-resolution provided by the default Edinburgh Geoparser are re-applied once the manual annotation (described in the next section) for a document is completed. This is to ensure that the corrected text mining output is geo-resolved and normalised correctly for dates, including manual corrections of spelling mistakes occurring in entity mentions.

The main effort in adapting the Edinburgh Geoparser was directed towards adding additional entity types to those recognised by the default version (e.g. geographic-feature, plague-ontology-term, and population). Plague related terminology (plague-ontology-term) is recognised using a domain-specific lexicon of terms relevant to the third plague pandemic. This was bootstrapped using manual annotation of plague terms in the pilot phase (including ones containing OCR errors) and extending this list by allowing matches of different forms (singular/plural) and crucially adding manually corrections of OCR errors as attributes to the entity annotations. This enables us to add annotations automatically to text still containing OCR errors with the aim to do further OCR post-correction or allow keyword searching over text containing these errors and including them in the results even if the search term is typed correctly. Geographic features are marked up using similar lexicon matching which is complemented by adding further geographical features that are derived automatically using WordNet,[20] a lexical database of semantic relations. The latter approach is also used to recognise population entities.

3.3 Manual Annotation

The bulk of the manual annotation has been carried out by two main annotators, a PhD student trained in natural language processing and a medical anthropologist PhD student. Prior to the main annotation phase we conducted a pilot annotation which lasted one week in order to train both annotators how to use the annotation tool, what information to mark up and to refine details in the annotation guidelines. This pilot annotation involved direct input and feedback from the academics leading this project (a computational linguist and a medical historian). After the pilot, the two annotators started annotating the data independently but asked any queries they had to the group. The balance of historians and NLP experts on this project worked as an advantage as the former bring the knowledge about the data, the historical background and ideas of what information is needing to be captured in the annotation and the latter have the expertise in the technology and methods used when applying natural language processing to automate or semi-automate some of the steps in this process.

Manual annotation was necessary for a number of reasons. Whilst some zones could be identified automatically from section titles we found that this was often hampered by spelling errors due to OCR issues arising from typeface styles and title placements in margins. In addition, depending on author narrative styles some zones could be found nested within sections with no titles. This created the need to manually annotate zones. The automatic recognition of named entities (see Section 3.2) was partially successful but also suffered from spelling errors and OCR issues. In addition, as reports were annotated new entity mentions were identified. Thus manual additions or correction of erroneous or spurious entity mentions was deemed necessary.

Manual annotation is conducted using Brat,[21] a web-based text annotation tool [Stenetorp et al.,

and ethnicity that were often implicated in the construction of epidemiological arguments. Some examples of the population/group of people annotations show that these are often derogatory and considered offensive today. They are strictly understood to be of value only for the illumination of historical discourse.

[20] https://wordnet.princeton.edu

[21] https://brat.nlplab.org/

2012]. After the text was processed automatically as described above, it was converted from XML into Brat format to be able to correct the text mining output and add zone annotations.[22] Figure 9 shows an excerpt of an example report being annotated in Brat. Entities such as date, location or geographic feature listed in Table 3 can be seen highlighted in the text. The start of an *outbreak history* zone is also marked at the beginning of the excerpt.

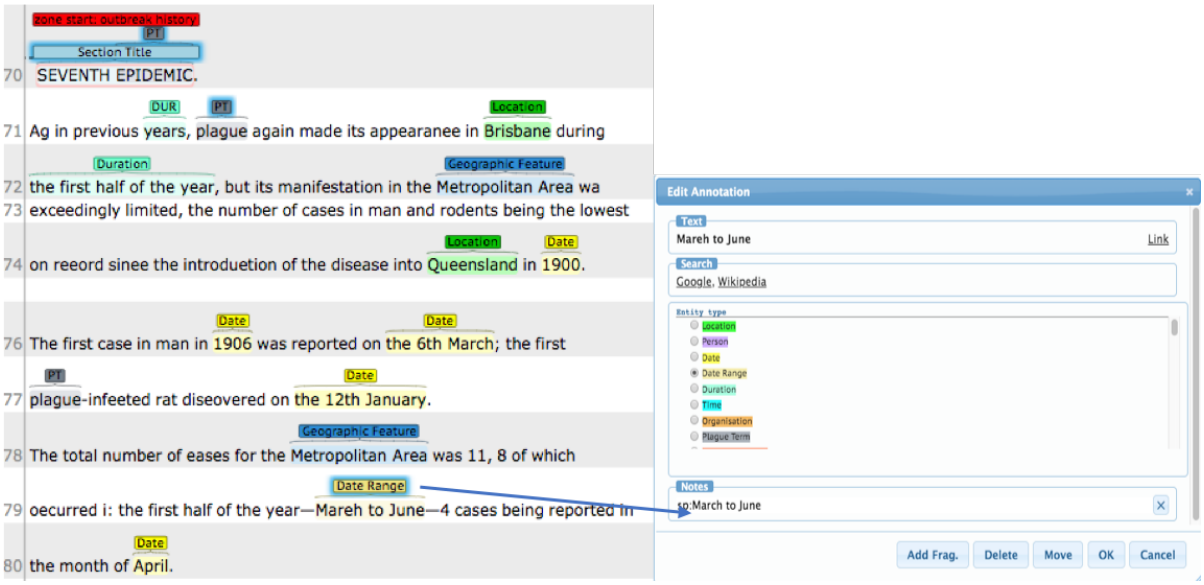


Figure 9: Brat annotation tool.

3.3.1 Zone Annotation

Zone annotation, as defined by our schema shown in Table 2, is applied inclusive of a section title and can be nested. For example, zones of *cases* are often found inside *treatment* or *clinical appearances* zones. *Footnote* zones were added as these often break the flow of the text and make downstream natural language processing challenging. In addition, we added *Header/Footer* markup to be able to exclude headers and footers, e.g. the publisher name or title or section title of a report repeated on each page, from further analysis or search.

Tables were a challenge for the OCR and unusable for the most part. When marking up tables, we also record their page number. Text within tables is currently ignored when ingesting the structured data to Solr (see Section IV). However, tables include a lot of valuable statistical information. In the next phase of the project we will investigate whether this information can be successfully extracted or whether it will need to be manually collated.

3.3.2 Entity Annotation

During manual annotation we instruct our annotators to correct any wrongly automated entities and add those that were missed. Any mis-spellings of entity mentions, mostly caused by the OCR process, are also corrected in the Note field in Brat. An example date-range annotation containing an OCR error, *Mareh to June* corrected to *March to June*, is shown in Figure 9. The mis-spellings are subsequently used as part of our text cleaning process. The corrected forms are also used to geo-resolve place names and normalise dates. These final two processing steps of the Edinburgh Geoparser are carried out on each report once it has been manually annotated and converted back to XML.

[22] We have not yet conducted double annotation to determine inter-annotator agreement for this work but this is something we are planning to do in the future.

IV DATA SEARCH INTERFACE

One goal of the *Plague.TXT* project is to make our digital collection available as an online search and retrieval resource but in addition this collection should be accessible. This means being available for example, to computational linguists as an annotated resource for direct in-depth analysis as well as via interactive search for humanities researchers. This provides vital support for historians and humanities researchers improving on the limited capacities of manual searches through document collections to find information pertinent to their research interest. Additionally, the challenges of working with such text digitally require interdisciplinary collaboration. HistSearch [Pettersson et al., 2016], an on-line tool applied to historical texts, demonstrates how computational linguists and historians can work together to automate access to information extraction and we will evaluate similar approaches for this collection. We plan to make the digital collection available with Apache Solr.[23]

Solr is an open-source enterprise-search platform, widely used for digital collections. The features available through the Solr search interface make our collection accessible to a wide audience with varying research interests. It offers features that support grouping and organising data in multiple ways, while data interrogation can be achieved through its simple interface with term, query, range and data faceting. Solr also supports rich document handling with text analytic features and direct access to data in a variety of formats.

We are currently customising and improving the filtering of the data for downstream analysis in Solr. In the following section, we describe on-going filtering steps with Solr and provide examples to demonstrate a search interface customisation. Further, we explore preliminary analysis that can be done from data retrieved via the search interface.

4.1 Data Preparation and Filtering in Solr

The annotated data is prepared and imported to Solr using Python, with annotations created both automatically by the Geoparser and manually by the annotators mapped to appropriate data fields (e.g. date-range entities are mapped to a Date Range field,[24]) enabling complex queries across the values expressed in the document text. Additionally, manual spelling corrections are used to replace the corresponding text in the OCR rendering prior to Solr ingestion, thus improving the accuracy of language-based queries and further textual analysis. We also implement lexicon-based entity recognition for entities that have been missed during the annotation and for additional entity types, e.g animals. Solr allows for storing and searching by geo-spatial coordinates and we import geo-coordinates associated with entities identified by the Geoparser. Geo-coordinates can be used to support interactive visualisations, as developed in the Trading Consequences project [Hinrichs et al., 2015] which visualises commodities through their geo-spatial history. In addition, this location information can be used in analysis such as transmission and spread, e.g. geo-referenced plague outbreak records have been used to show how major trade routes contributed to the spread of the plague [Yue et al., 2017]. Using *case* zones we are currently assessing NLP techniques to extract case information into a more structured format for direct access to statistical information from hundreds of individual case descriptions.

V USE CASES FOR THE PLAGUE.TXT DATA

Historians and computational linguists have different methods and reasons for analysing a data set. The Plague.TXT team not only provide a search and exploration interface but will also

[23] https://lucene.apache.org/solr/
See this website for a further description of Solr features.

[24] https://lucene.apache.org/solr/guide/8_1/working-with-dates.html#date-range-formatting

release the underlying data (for titles with permissible licenses) to allow direct corpus analysis. In this section, we provide examples for three different use cases of this data.

5.1 Use Case 1: Illustration of Interactive Search

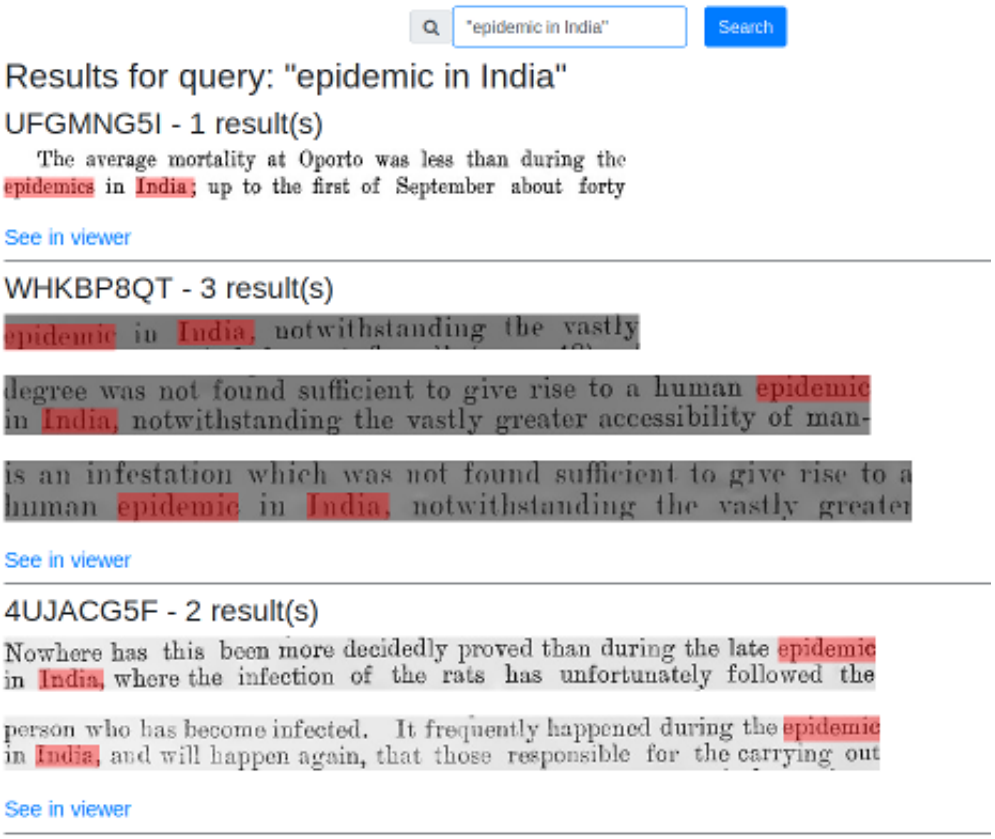


Figure 10: Solr snippet search example.

Figure 10 shows one of our customised search interfaces. This allows users to search across the entire report collection displaying original image snippets from the reports containing the search term(s). This search function enables the user to grasp the immediate context of search-terms within the page and also to recognise potential limits of the OCR recognition currently applied.

Snippet search is supported by indexing OCR transcriptions from word-level ALTO-XML[25] in Solr and then by using Whiiif,[26] an implementation of the International Image Interoperability Framework (IIIF) Search API[27] designed to provide full-text search with granular, word-level annotation results to enable front-end highlighting.

Figure 11 shows a similar search within a single document using UniversalViewer,[28] a IIIF viewer utility. This document-level search is powered by the same Whiiif[29] instance, again making use of the IIIF Search API to provide a method of in-document searching that is available natively within any compatible IIIF viewing software. This functionality is made available

[25] http://www.loc.gov/standards/alto/
[26] https://github.com/mbennett-uoe/whiiif
[27] https://iiif.io/api/search/1.0/
[28] https://universalviewer.io
[29] Whiiif stands for Word Highlighting for IIIF. Further technical details about the Whiiif package can be found on the University of Edinburgh Library Labs blog: http://libraryblogs.is.ed.ac.uk/librarylabs/2019/07/03/introducing-whiiif/

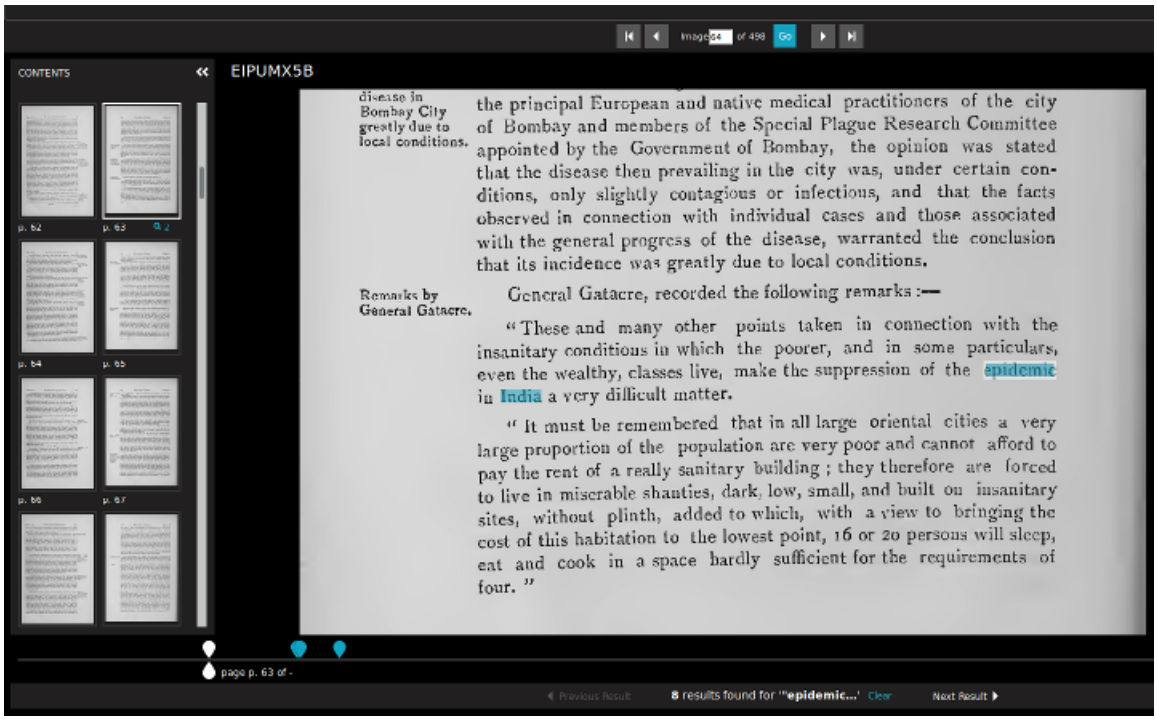


Figure 11: IIIF viewer search example.

to visualise the search results within the document context, as part of a whole-document browsing interface, allowing for greater context of the search results to be shown.

5.2 Use Case 2: Finding Discussion Concepts in Causes Across Time Periods

Our search interfaces facilitate queries across the collection on content and meta-data as well as queries based on zones or entity types using facets such as date range. In this use case we focus on topics discussed in *cause* zones and if these differ between report time periods. We do this through extracting the data via Solr and applying topic modelling. First, we search for *cause* zones published during the pandemic 1894-6 comparing these to *cause* zones in reports 1904 and beyond. We retrieve the results via the Solr API in XML format and apply removal of stop-words and all non dictionary terms to the *cause* zone text. In future, we will make indexed versions of cleaned data in this format directly accessible from Solr. We use topic modelling (LDA with Gensim Python library[30]) to compare the *cause* zone text at the different time points, selecting two topics.

Results are presented in Table 4. The earlier reports show the discussion centering around environment aspects with focus on populations, conditions of living and buildings and how this might cause the spread. The second topic is linked to the concepts at that time period, about how the diseases may spread through the water system, with studies of ordinance maps of sewerage and water ways. Looking at the later reports we now see rats and fleas and infection are more prominent as a discussion topic but also season, temperature and weather form a topic being discussed as a causal factor. Combined further with geo-resolution information, this type of zone-specific topic analysis across time periods could reveal interesting patterns and inferences about the reasoning of epidemiologists observing outbreaks.

[30] https://radimrehurek.com/gensim/index.html

Topic/Date	keywords
(1) 1894-96	latrine, house, soil, street, find, case, time, plague, infection, opinion, condition, may, must, question, see
(2) 1894-96	house, people, ordinance, well, supply, cause, must, condition, drain, disease, pig, matter, area, water, provision
(1) 1904-07	plague, rat, case, infection, man, flea, may, infect, place, fact, evidence, disease, instance, produce, find
(2) 1904-07	year, month, temperature, epidemic, influence, season, infection, december, condition, may, june, prevalence, rat, follow, number

Table 4: Discussion topics from cause zones by time period

5.3 Use Case 3: Corpus Analytics

Analysing a corpus with respect to token frequencies can reveal interesting patterns and insights into aspects such as gender, age and population. In this use case we look at the corpus with respect to gender and consider how men and women are mentioned within the corpus. As all reports in the collection are tokenised and part-of-speech tagged, frequency-based and syntactic-category-dependent corpus analysis can be conducted across the collection.[31] The ratio of the total number of mentions of *woman* or *women* versus *man* or *men* is 0.19 (681 versus 3603 mentions after lower-casing the text). Similarly the ratio for the pronouns *she* versus *he* is 0.15 (1233 versus 8008 mentions after lower-casing).

Table 5 lists the twenty most frequent adjectives followed by *man/men* versus *woman/women*. For the majority of mentions, men are described as *medical*, *young* and *sick* and women as *old*, *married* and *pregnant*. A similar analysis for verbs following pronouns, the most frequent verbs following *he* (excluding *has*, *is*, *could*, *would* etc.) are *thought* (n=119), *died* (n=80) and *found* (n=65). The phrase *she died*, on the other hand, appears only 15 times out of over 4.2 million words in the collection. This difference is comparable with the ratio of mentions of woman/woman versus man/men (0.19). However, the ratio is much more skewed for the phrases *she thinks/thought* (n=2) versus *he thinks/thought* (n=144).

While these results are unsurprising given that the reports were written over a century ago and authored by men for men, they do raise questions on gender statics within the reporting of cases in these reports. More thorough analysis, for example by exploring this text in context of its time (see the DICT method proposed by Jatowt et al. [2019]), combined with close reading is necessary to explore these differences in more detail. The search interface to the collection,

[31] A syntactic category corresponds to a part of speech of a text token (e.g. noun, verb, preposition, etc.). Syntactic-category dependent corpus analysis is therefore counting tokens that are tagged with a particular part of speech tag.

adjective + man|men		adjective + woman|women	
count	adjective	count	adjective
316	medical	21	old
27	young	12	married
22	sick	10	pregnant
19	old	9	young
13	medieal	8	chinese
9	poor	7	purdah
8	influential	5	native
7	other	4	other
7	infected	3	parturient
7	healthy	3	dead
6	intelligent	2	well-nourished
6	few	2	unfortunate
5	well-nourished	2	indian
5	twelve	2	few
5	scientific	1	weakly
4	white	1	sick
4	trained	1	several
4	several	1	russian
4	muscular	1	respectable
4	great	1	purdak

Table 5: Most frequent adjectives followed by the nouns *man* or *men* versus *woman* or *women*.

however, helps to find instances of these mentions in, for example, the context of *case* zones, and thereby supports navigation of the collection more rapidly.

DISCUSSION, CONCLUSIONS AND FUTURE WORK

In this paper we have presented the work undertaken in the pilot stage of our *Plague.TXT* project. The work is the outcome of an interdisciplinary team working together to understand the nature and complexities of a historical text collection and the needs of the potential different types of users of this collection. A major contribution of this project is the development of a model to capture the narrative structure of the collection of reports. This brings individual reports together in one collection enabling streamlined and efficient linking of knowledge and themes used in the comprehension of the third plague pandemic, covering the time period of the collection. This approach enables analysis of these reports across sections as one coherent corpus. Making this collection accessible through the Solr search interface, we can share it with the research community in ways that cater for the needs of different field experts

Our work in this project is ongoing as we add more data but manual annotation is time consuming and can be an error prone process. As we increase the number of reports annotated with zone markup, we intend to investigate how the annotation can be automated. Possible solutions include: methods, such as content similarity measures which have been shown to be successful in scientific article recommendation [He et al., 2010], or work in identifying clinical note duplication [Wrenn et al., 2010] which uses distance between words, or work that measures similarity of scientific articles using divergence of distributions of words [Huang et al., 2019].

Currently we are developing methods to directly access the statistical information contained within *case* zones and within tables. Additionally, we will explore spelling normalisation fur-

ther, such as diachronic and synchronic spelling variance. As well as the methods for spelling normalisation previously mentioned we will also use fuzzy string matching capabilities within Solr to correct for spelling variation introduced by OCR. Additionally, we will explore changes in semantics over time and how this may impact search and downstream analysis.

ACKNOWLEDGEMENTS

This work was funded by the Challenge Investment Fund 2018-19 from the College of Arts, Humanities and Social Sciences, University of Edinburgh.

References

B. Alex and J. Burns. Estimating and rating the quality of optically character recognised text. In *Proceedings of DATeCH 2014*, pages 97–102, 2014. URL http://doi.acm.org/10.1145/2595188.2595214.

B. Alex, C. Grover, E. Klein, and R. Tobin. Digitised Historical Text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 401–409, 2012. URL http://www.oegai.at/konvens2012/proceedings/59_alex12w/59_alex12w.pdf.

B. Alex, K. Byrne, C. Grover, and R. Tobin. Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal for Humanities and Arts Computing*, 9(1):15–35, 2015. URL https://www.euppublishing.com/doi/abs/10.3366/ijhac.2015.0136.

B. Alex, C. Grover, R. Tobin, and J. Oberlander. Geoparsing historical and contemporary literary text set in the City of Edinburgh. *Language Resources and Evaluation*, 53(4):651–675, 2019. URL https://doi.org/10.1007/s10579-019-09443-x.

V.K. Bhatia. Analysing Genre: Language use in Professional Settings. *New Youk:Routledge*, 2014.

Marcel Bollmann. automatic normalization of historical texts using distance measures and the norma tool. In *Proceedings of the second workshop on annotation of corpora for research in the humanities (ACRH-2), Lisbon, Portugal*, pages 3–14, 2012.

Marcel Bollmann, Joachim Bingel, and Anders Søgaard. Learning attention for historical text normalization by learning to pronounce. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 332–344, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1031. URL https://www.aclweb.org/anthology/P17-1031.

B. Bramanti, K.R. Dean, L. Walløe, and N.C. Stenseth. The third plague pandemic in europe. *Proceedings of the Royal Society B: Biological Sciences*, 286(1901):20182429, 2019. doi: 10.1098/rspb.2018.2429. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2018.2429.

Lyle Campbell. *Historical linguistics*. Edinburgh University Press, 2013.

J. Clifford, B. Alex, C.M. Coates, E. Klein, and A. Watson. Geoparsing history: Locating commodities in ten million pages of nineteenth-century sources. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49(3):115–131, 2016. doi: 10.1080/01615440.2015.1116419. URL https://doi.org/10.1080/01615440.2015.1116419.

K.R. Dean, F. Krauer, and B.V. Schmid. Epidemiology of a bubonic plague outbreak in Glasgow, Scotland in 1900. *Royal Society Open Science*, 6(1):181695, 2019. doi: 10.1098/rsos.181695. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsos.181695.

M. J. Echenberg. *Plague Ports: The Global Urban Impact of Bubonic Plague, 1894-1901*. New York University Press, New York, 2007. ISBN 978-0-8147-2232-9.

L. Engelmann. Mapping Early Epidemiology: Concepts of Causality in Reports of the Third Plague Pandemic 1894–1950. In E. T. Ewing and K. Randall, editors, *Viral Networks: Connecting Digital Humanities and Medical History*, pages 89–118. VT Publishing, 2018. ISBN 978-1-949373-02-8. doi: https://publishing.vt.edu/site/books/10.21061/viral-networks/. URL https://vtechworks.lib.vt.edu/handle/10919/86368.

B. Fu, M. Wu, R. Li, W. Li, Z. Xu, and C. Yang. A model-based book dewarping method using text line detection. In *Proc. CBDAR 2007*, pages 63–70, 2007. URL http://imlab.jp/cbdar2007/proceedings/papers/P1.pd.

A. Gotscharek, U. Reffle, C. Ringlstetter, K. U. Schulz, and A. Neumann. Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *IJDAR*, 14(2):159–171, 2011. URL https://link.springer.com/article/10.1007/s10032-010-0132-6.

C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball. Use of the Edinburgh Geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A*, 368(1925):

3875–3889, 2010. URL https://doi.org/10.1098/rsta.2010.0149.

Claire Grover and Richard Tobin. Rule-based chunking and reusability. In *Proceedings of LREC 2006*, pages 873–878, 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/457_pdf.pdf.

A. Hauser, M. Heller, E. Leiss, K. U. Schulz, and C. Wanzeck. Information access to historical documents from the Early New High German period. In L. Burnard, M. Dobreva, N. Fuhr, and A. Lüdeling, editors, *Digital Historical Corpora- Architecture, Annotation, and Retrieval*, Dagstuhl, Germany, 2007. URL https://drops.dagstuhl.de/opus/volltexte/2007/1057/.

Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 421–430, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772734. URL http://doi.acm.org/10.1145/1772690.1772734.

Marcin Heliński, Miłosz Kmieciak, and Tomasz Parkoła. Report on the comparison of tesseract and abbyy finereader ocr engines. 2012. URL https://www.digitisation.eu/fileadmin/Tool_Training_Materials/Abbyy/PSNC_Tesseract-FineReader-report.pdf.

U. Hinrichs, B. Alex, J. Clifford, A. Watson, A. Quigley, E. Klein, and C.M. Coates. Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*, 30(suppl_1):i50–i75, 10 2015. ISSN 2055-7671. doi: 10.1093/llc/fqv046. URL https://doi.org/10.1093/llc/fqv046.

Chien-yu Huang, Arlene Casey, Dorota Głowacka, and Alan Medlar. Holes in the outline: Subject-dependent abstract quality and its implications for scientific literature search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 289–293, 2019. URL https://dl.acm.org/doi/10.1145/3295750.3298953.

Adam Jatowt, Ricardo Campos, Sourav S. Bhowmick, and Antoine Doucet. Document in context of its time (dict): Providing temporal context to support analysis of past documents. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2869–2872, 2019. doi: 10.1145/3357384.3357844. URL https://doi.org/10.1145/3357384.3357844.

E. Klein, B. Alex, and J. Clifford. Bootstrapping a historical commodities lexicon with SKOS and DBpedia. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 13–21, 2014. URL https://www.aclweb.org/anthology/W14-0603.pdf.

Natalia Korchagina. Normalizing medieval German texts: from rules to deep learning. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 12–17, Gothenburg, May 2017. Linköping University Electronic Press. URL https://www.aclweb.org/anthology/W17-0504.

M Liakata, S Saha, S Dobnik, and C Batchelor. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000, 2012. URL https://www.ncbi.nlm.nih.gov/pubmed/22321698.

D. Lopresti. Measuring the impact of character recognition errors on downstream text analysis. In B.A. Yanikoglu and K. Berkner, editors, *Document Recognition and Retrieval*, volume 6815. SPIE, 2008. URL https://doi.org/10.1117/12.767131.

J.A. Lowson. *The Epidemic of Bubonic Plague in Hongkong, 1894*. Noronha & Company, Hong Kong, 1895. URL https://archive.org/details/b24398287.

Petar Mitankin, Stefan Gerdjikov, and Stoyan Mihov. An approach to unsupervised historical text normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, page 29–34, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325882. doi: 10.1145/2595188.2595191. URL https://doi.org/10.1145/2595188.2595191.

A. Morabia, editor. *A history of epidemiologic methods and concepts*. Birkhauser Verlag, Basel ; Boston, 2004. ISBN 978-3-7643-6818-0. OCLC: 55534998.

M.S. Morgan and M.N. Wise. Narrative science and narrative knowing. Introduction to special issue on narrative science. *Studies in History and Philosophy of Science Part A*, 62:1–5, 2017. ISSN 00393681. doi: 10.1016/j.shpsa.2017.03.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0039368117300729.

E. Pettersson, J. Lindström, B. Jacobsson, and R. Fiebranz. Histsearch – implementation and evaluation of a web-based tool for automatic information extraction from historical text. In *3rd HistoInformatics Workshop*, Krakow, Poland, 2016. URL http://ceur-ws.org/Vol-1632/paper_4.pdf.

Eva Pettersson. *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. PhD thesis, Uppsala University, Department of Linguistics and Philology, 2016.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 163–179, Oslo, Norway, May 2013a. Linköping University

Electronic Press, Sweden. URL https://www.aclweb.org/anthology/W13-5617.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. An smt approach to automatic annotation of historical text. *Workshop on Computational Historical Linguistics, Nodalida 2013*, 01 2013b.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0605. URL https://www.aclweb.org/anthology/W14-0605.

Michael Piotrowski. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157, 2012.

Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. A deeply annotated testbed for geographical text analysis: the corpus of Lake District writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 9–15. ACM, 2017. URL https://dl.acm.org/doi/10.1145/3149858.3149865.

Don Ringe and Joseph Eska. *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press, 2013.

C.J. Rupp, P. Rayson, A. Baron, C. Donaldson, I. Gregory, A. Hardie, and P. Murrieta-Flores. Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data*, pages 59–62. IEEE, 2013. URL https://doi.org/10.1109/BigData.2013.6691671.

Yves Scherrer and Tomaž Erjavec. Modernizing historical slovene words with character-based smt. pages 58–62, 08 2013.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of EACL 2012*, pages 102–107, 2012. URL https://www.aclweb.org/anthology/E12-2021/.

J.M. Swales. *Genre Analysis: English in academic and research settings*. Cambridge University Press, 1990.

S. Teufel. *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh, 1999.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D09-1155.

Paul Thompson, Riza Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou. Text mining the history of medicine. *PloS one*, 11:e0144717, 01 2016. doi: 10.1371/journal.pone.0144717.

Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53, 01 2010. ISSN 1067-5027. doi: 10.1197/jamia.M3390. URL https://doi.org/10.1197/jamia.M3390.

A. Yersin. La Peste Bubonique a Hong Kong. *Annales de l'Institut Pasteur*, pages 662–667, 1894. f667, 1.

R. Yue, H.F. Lee, and C.Y.H. Wu. Trade routes and plague transmission in pre-industrial Europe. *Scientific reports*, 7(1):12973, 2017. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=29021541&retmode=ref&cmd=prlinks.